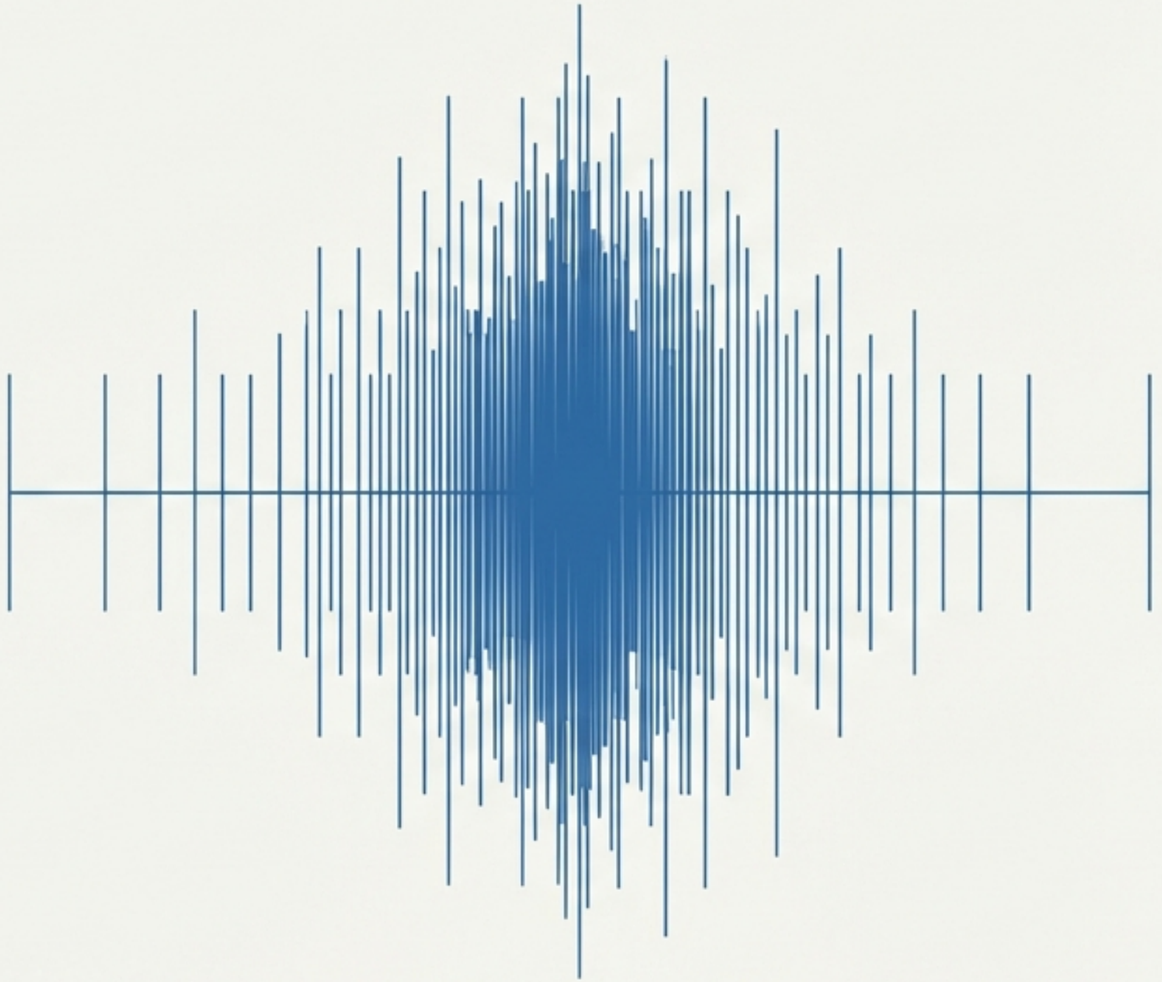
Statistical Dispersion: The Science of Variability

Chapter 5: Measures of Variability



This presentation covers the essential measures of spread within a dataset, moving from basic estimates like Range to precise calculations like Standard Deviation.

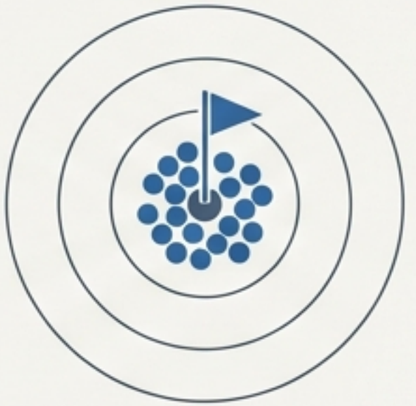
NotebookLM

Measures of Variability – Consistency in Performance

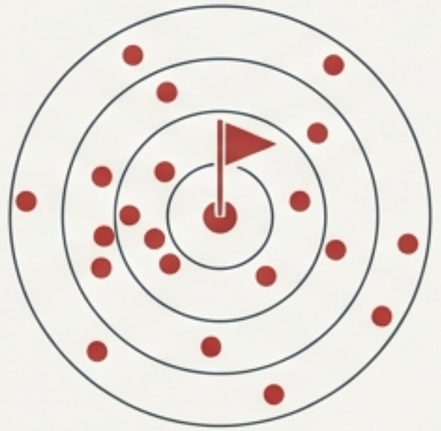
Consistency is often as critical as ability; consider two golfers who both hit 25 balls with an average distance from the pin of zero. The professional is consistent because her shots are clustered closely around the pin, representing low variability and high precision.
The beginner is inconsistent because his shots scatter widely to the left, right, long, and short, creating high variability even though the mathematical average of his errors balances out to zero.

If we only evaluated the mean, we would erroneously conclude they are equal in ability, which highlights why measuring the spread or dispersion of data is necessary to understand performance.

Professional



Beginner



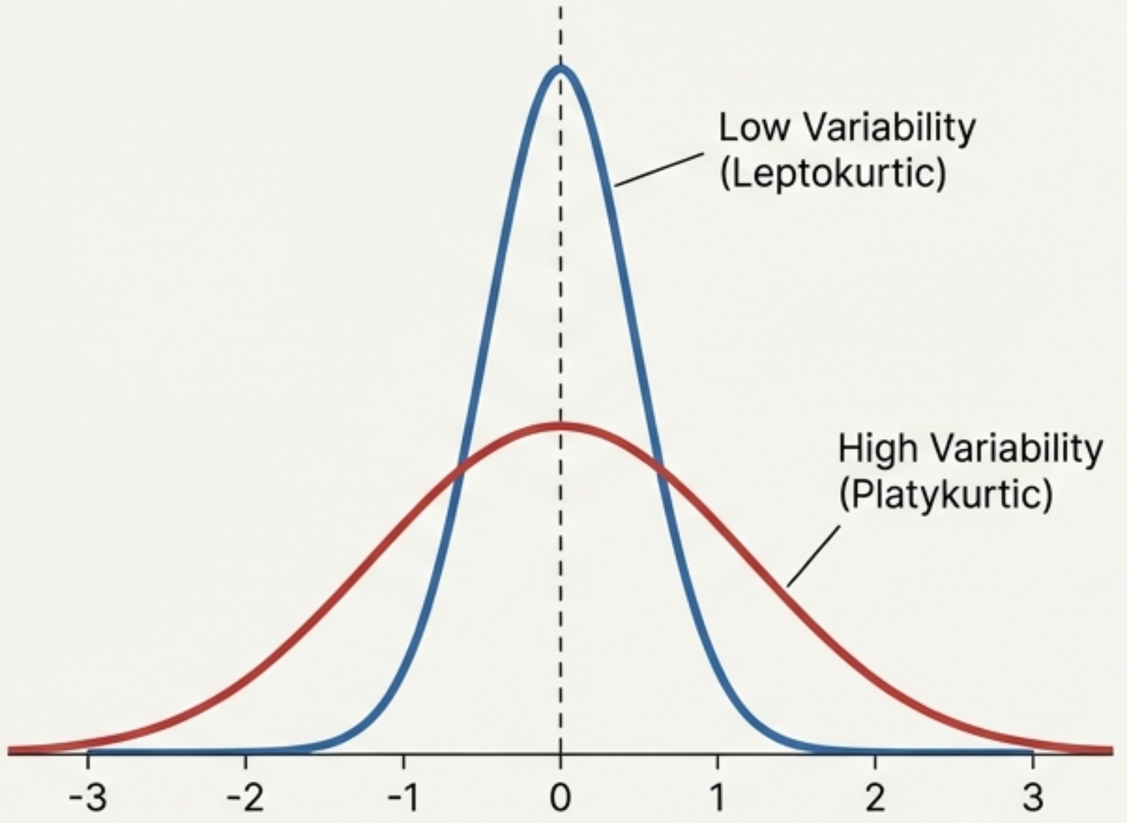
NotebookLM

Measures of Variability – Means vs. Dispersion

Variability is the quantitative measure of the spread or dispersion of a set of data. Data that is compact around the central measure is represented by leptokurtic curves, while data that is widely spread out is represented by platykurtic curves.

It is often interesting to compare the means of two datasets to see which is higher, but it is equally important to compare their variabilities because two groups can have identical means but completely different spreads.

Statisticians rely on four main measures to quantify this spread: the range, the interquartile range, the variance, and the standard deviation.



NotebookLM

Range - Calculation and Limitations

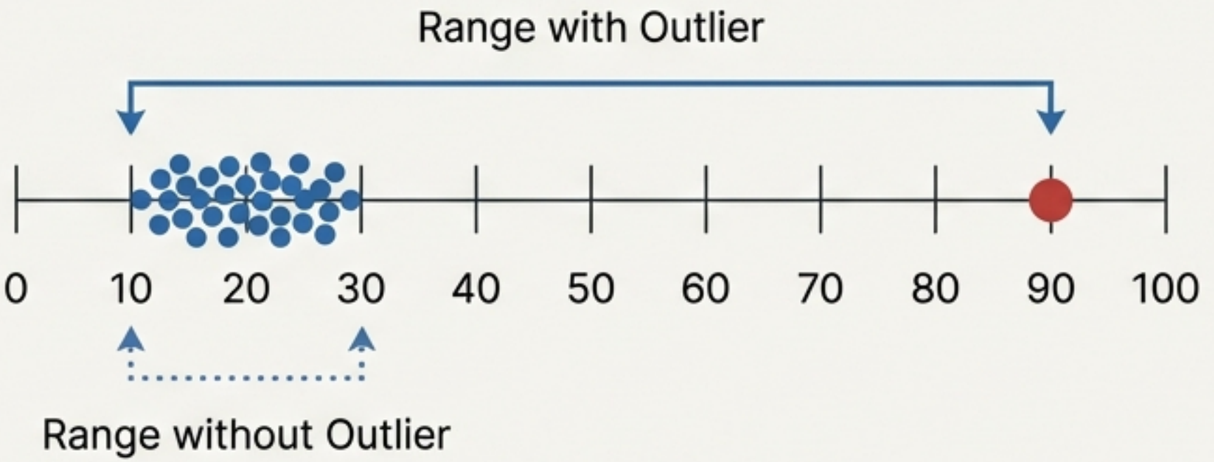
The range is the simplest measure of variability and is calculated as the difference between the highest and lowest scores in a data set.

$$\text{Range} = \text{High Score} - \text{Low Score}$$

The “Outlier Problem”

While it provides the quickest estimate of variability, the range is considered a very unstable measure because it relies entirely on the two most extreme scores.

Because it ignores all the data points between the extremes, the introduction of a single outlier score can radically change the range, making it a rough estimator rather than a precise metric.



NotebookLM

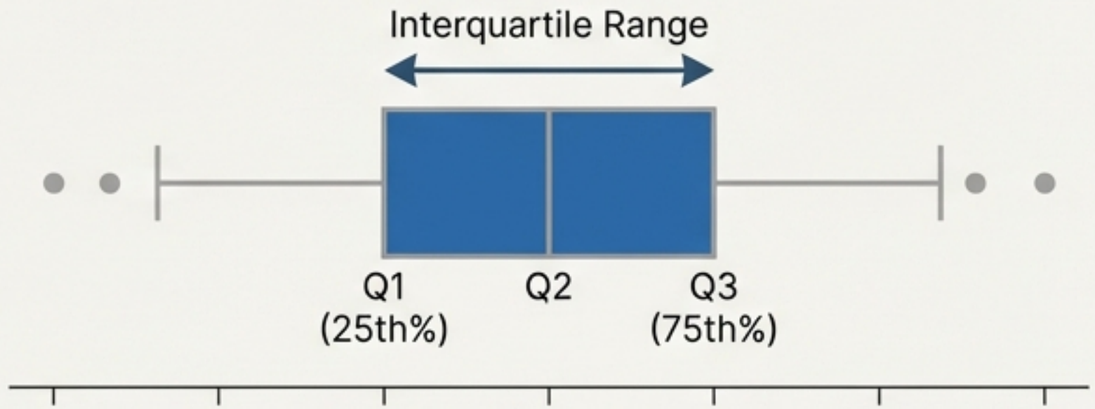
Interquartile Range - Ordinal and Skewed Data

The Interquartile Range (IQR) is calculated by finding the difference between the raw scores at the 75th percentile (Q3) and the 25th percentile (Q1). This measure is particularly useful for ordinal data or highly skewed interval/ratio data because it focuses exclusively on the middle 50% of the scores, effectively ignoring highly divergent outliers at the extremes.

For situations requiring a smaller indicator of variability, researchers may use the Semi-Interquartile Range (SIQR), which is simply half of the Interquartile Range.

$$\text{IQR} = Q_3 - Q_1$$

$$\text{SIQR} = \frac{Q_3 - Q_1}{2}$$



NotebookLM

Variance - Deviations from the Mean in Playfair Display

To assess variability more precisely than the range or IQR, we must determine the distance, or deviation, of every single raw score from the mean of the data.

However, a mathematical constraint exists: the sum of the deviations around the mean will always equal zero because the negative deviations below the mean perfectly balance the positive deviations above it. If the deviations do not sum to zero, it indicates that the mean was calculated incorrectly, but this zero-sum property prevents us from simply adding the deviations to find the total variability.

Table 5.1 Deviations From the Mean

X (cm)	(X - Mean)
27	+2
26	+1
25	0
24	-1
23	-2
Sum = 125	Sum of deviations = **0**

NotebookLM

Variance - Calculation and Logic

While we could sum the absolute values of deviations, this is algebraically problematic for advanced statistics; instead, we eliminate negative signs by squaring each deviation.

Variance is formally defined as the average of the squared deviations from the mean. By squaring the deviations, all values become positive, allowing us to calculate a meaningful average that represents the spread of the data, though the result is now expressed in 'squared units' rather than the original units.

Table 5.2 Calculation of Variance

X (cm)	(X - Mean)	(X - Mean)²
27	+2	4
26	+1	1
25	0	0
24	-1	1
23	-2	4
Sum	Sum of Squares = **10**	

Variance = 10 / 5 = **2 cm²**

NotebookLM

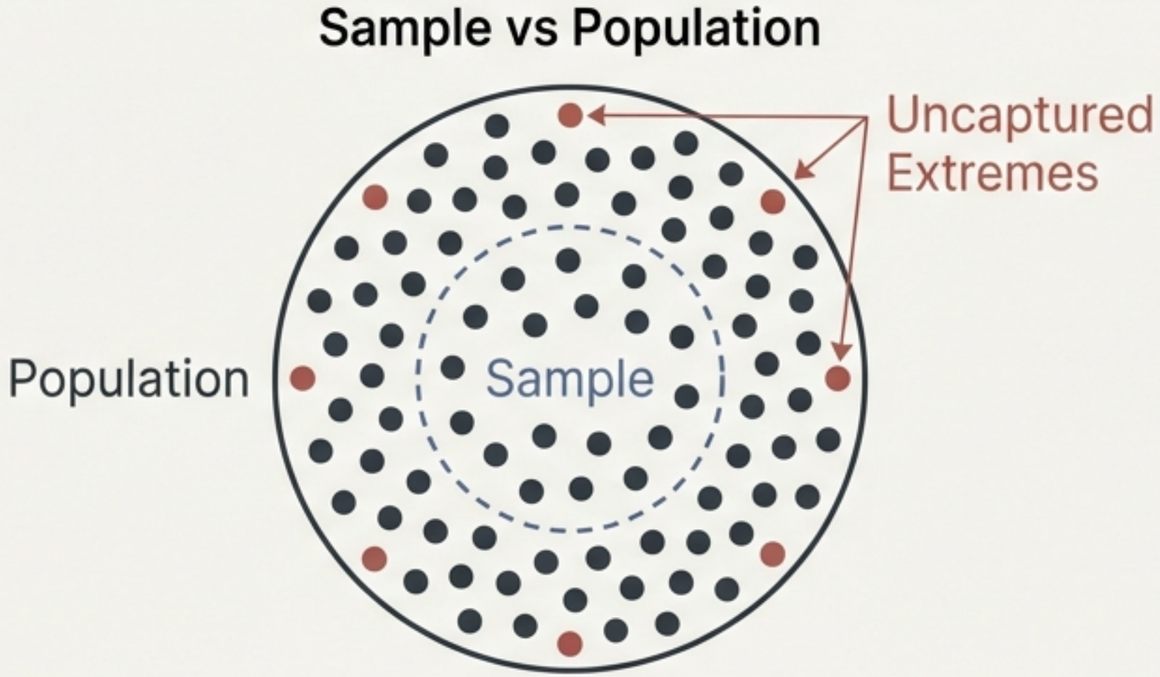
Standard Deviation - Population Definition

Because Variance is calculated in squared units (like ‘squared centimeters’), it is often difficult to interpret in the context of the original data. To return to the original unit of measurement, we calculate the Standard Deviation, which is simply the square root of the Variance. This statistic, symbolized by the Greek lowercase sigma (σ) for populations, provides an accurate description of variability that considers every data point, making it the foundation for many advanced statistical techniques.

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

NotebookLM

Calculating Standard Deviation for a Sample



Research is rarely conducted on entire populations; instead, samples are taken with the assumption that they represent the population. However, samples rarely contain the extreme values found in the full population (for example, the very heaviest or lightest individuals), meaning a sample's variability is almost always smaller than the true population variability.

To avoid underestimating the population's standard deviation, we must apply a mathematical correction factor to the sample calculation.

NotebookLM

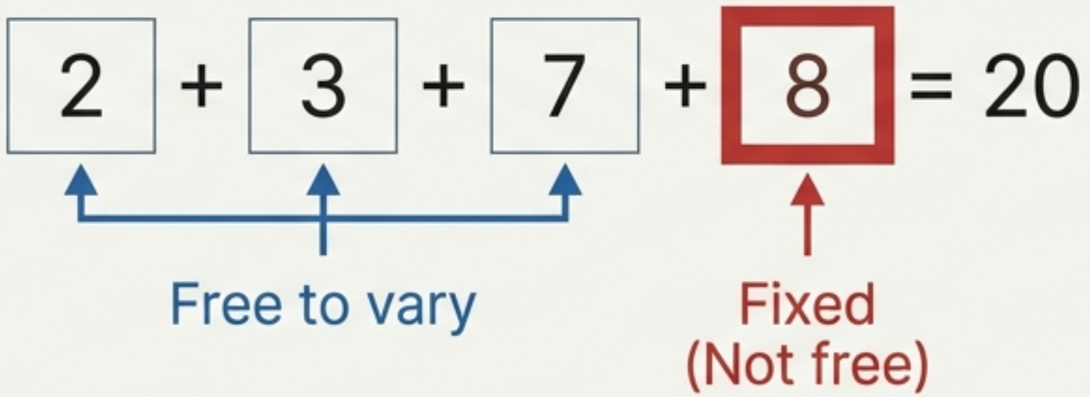
Standard Deviation - Degrees of Freedom

The correction for sample bias is based on the concept of degrees of freedom, calculated as N − 1.

Consider a sample of four numbers that must sum to 20 to maintain a specific mean; the first three numbers can vary freely (e.g., 2, 3, and 7), but the fourth number *must* be 8 to equal 20, meaning it is not free to vary.

In the formula for Sample Standard Deviation (SD), we divide the sum of squared deviations by N − 1 rather than N, which results in a slightly larger and more accurate estimate of the population's variability.

$$SD = \sqrt{\frac{\sum(X - \bar{X})^2}{N - 1}}$$



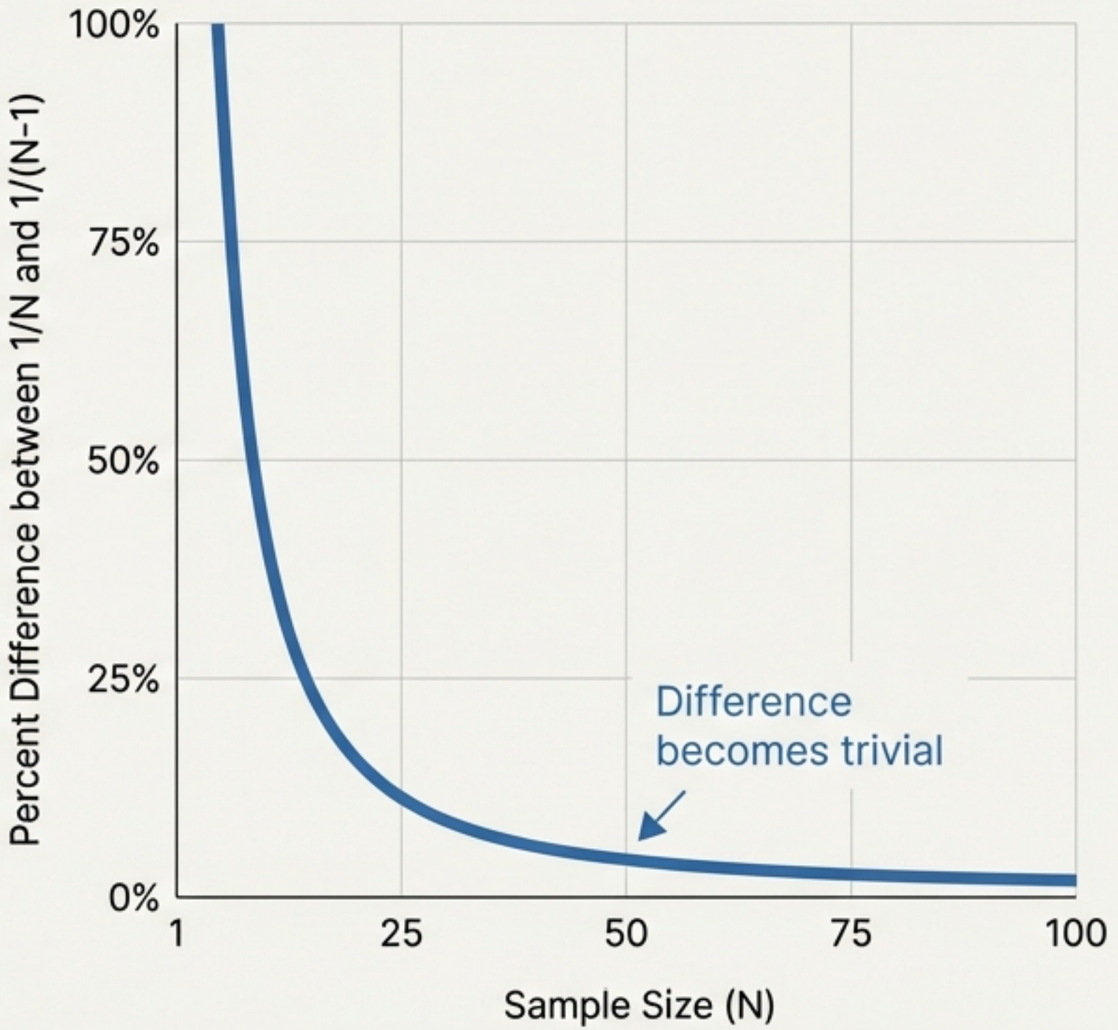
NotebookLM

Standard Deviation – Sample Size Accuracy

The difference between the population formula (dividing by N) and the sample formula (dividing by N - 1) is significant when the sample size is small. As the

sample size (N) increases, specifically above 50, the mathematical difference between dividing by N and N - 1 becomes trivial.

This mathematical behavior demonstrates that samples with a large N are inherently **more accurate** in estimating population parameters than samples with a small N.



NotebookLM

Coefficient of Variation - Comparing Variables

Researchers often need to compare the variability of two completely different variables, such as strength measured in kilograms versus vertical jump measured in centimeters. Because the units differ, a direct comparison of standard deviations is impossible; instead, we use the Coefficient of Variation (CoV).

The CoV normalizes the standard deviation by dividing it by the mean and multiplying by 100, resulting in a percentage that allows for a direct comparison of relative variability across different metrics.

$$\text{CoV} = \frac{\text{SD}}{\text{Mean}} \times 100$$

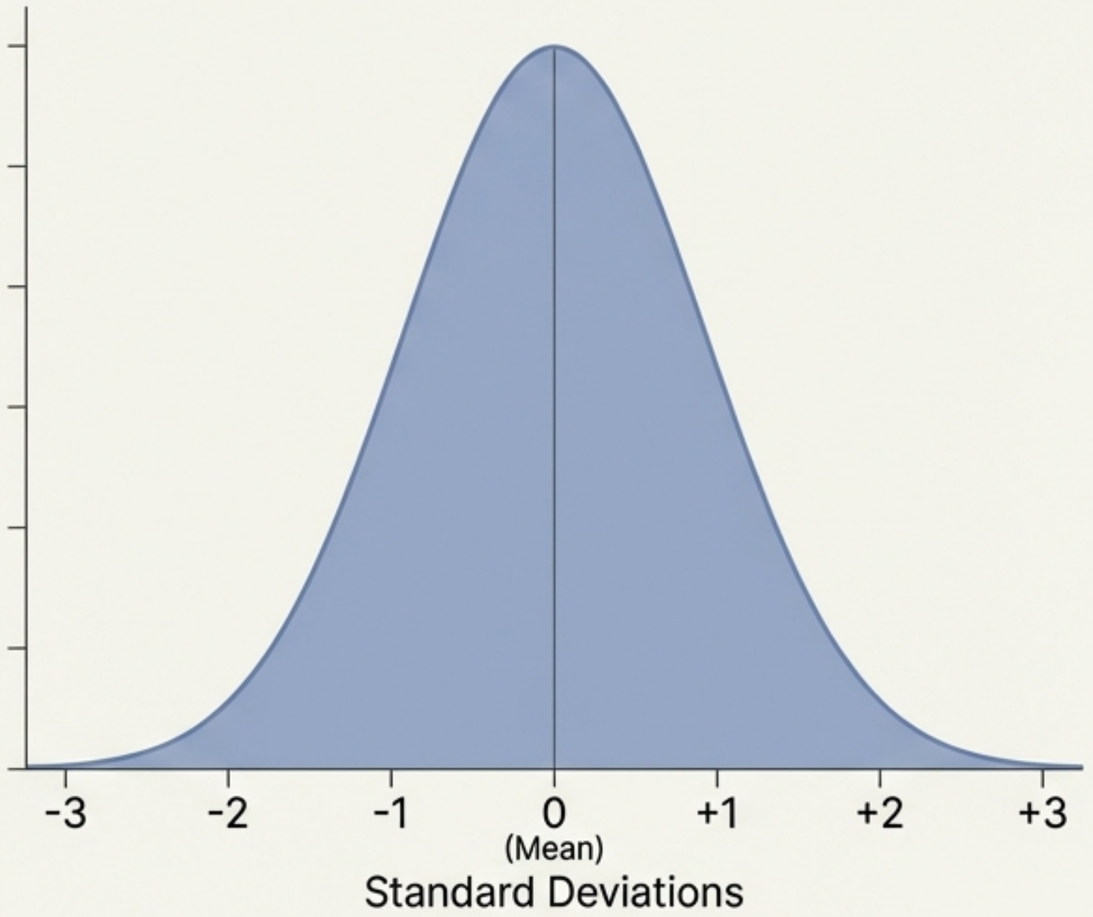
Variable A (Strength)	Variable B (Jump)
Mean = 100kg	Mean = 50cm
SD = 10kg	SD = 10cm
CoV = 10%	CoV = 20%

Conclusion: Variable B is relatively more variable.

NotebookLM

Standard Deviation – Normal Distribution

A useful rule of thumb for large, normally distributed datasets is that the range usually spans about 5 or 6 standard deviations. Specifically, there are typically about 3 standard deviations from the mean to the largest score and 3 standard deviations from the mean to the smallest score. While this rule does not hold for skewed data or small sample sizes, it allows a researcher to quickly check calculations for gross errors when analyzing large datasets.



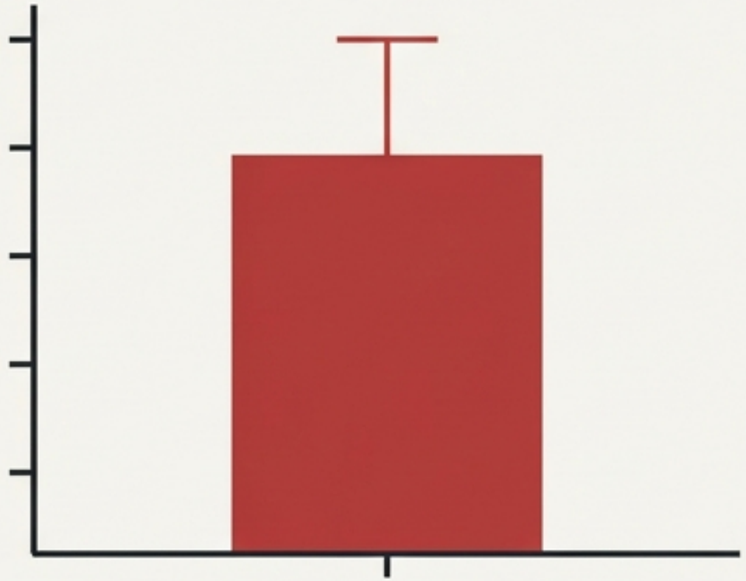
NotebookLM

Graphical Depiction – Reporting and Visuals

When reporting results, researchers typically denote variability in the text as 'Mean ± SD' (e.g., 25 ± 1.6 cm). Graphically, **"Dynamite plunger plots" (bar graphs with error bars)** are common but often criticized for **hiding the distribution of the underlying data. Dot plots** are increasingly preferred because they display the individual data points alongside the **mean** and **standard deviation**, revealing **skewness** or **outliers** that a simple bar graph would conceal.

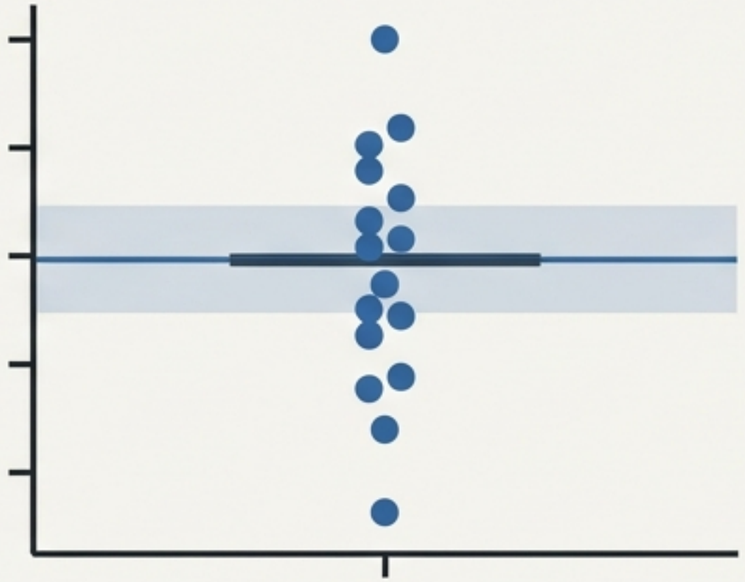
The Bad vs. The Good

Dynamite Plunger Plot



Conceals Data

Dot Plot



Reveals Distribution

NotebookLM

Summary and Activities

While the Range and IQR provide quick estimates, the Variance and Standard Deviation are the essential tools for analyzing dispersion in kinesiology research. Understanding these calculations is critical because they form the basis for determining relationships between variables (correlation) and comparing group differences (t-tests) in future chapters. Students should utilize the web resource activities provided to practice these calculations, specifically focusing on the difference between sample and population formulas.

Scan for Web Resource Activities

NotebookLM